Factores de influencia sobre la frecuencia de accidentes en la red de carreteras del estado, mediante árboles de regresión y clasificación.

Blanca Arenas Ramírez[(1)]
Francisco Aparicio Izquierdo[(1)]
Camino González Fernández[(2)]
[(1)]INSIA. Escuela Técnica Superior de Ingenieros Industriales. UPM. [(2)]Laboratorio de Estadística. Escuela Técnica Superior de Ingenieros Industriales. UPM.

RESUMEN

En este trabajo se ha explorado la aplicación de modelos de árboles de regresión y clasificación para el comportamiento de la variable dependiente accidentes, con la intensidad de tráfico, el porcentaje de vehículos pesados, una medida de la velocidad y el tipo funcional de carretera en tramos de la red de carreteras del Estado. Este tipo de modelos de minería de datos permiten extraer patrones de comportamiento e interacciones no fácilmente detectables mediante otro tipo de análisis.

Estos modelos de gran potencial, permitieron extraer información relevante, en lo concerniente a:

- La relevancia de las variables disponibles para la explicación del fenómeno de los accidentes.
- Las interacciones complejas existentes.
- Patrones de comportamiento.

Mediante los modelos de árboles de regresión para la variable respuesta *número total de accidentes* se ha evaluado el grado de importancia de las variables explicativas disponibles, así como las interacciones existentes entre las distintas variables. La variable de mayor influencia para la explicación de la frecuencia de accidentes en las carreteras es el tráfico medio anual, medida indirecta de la exposición en un tramo concreto. A medida que aumenta el tráfico total, mayor es el número medio de accidentes. El segundo factor que explica el número de accidentes es el porcentaje de vehículos pesados y también aumenta con el mismo. El tipo de vía, también es un factor de influencia sobre la frecuencia: en la muestra seleccionada hay mayor número de accidentes en autopistas y vías convencionales.

Los modelos de árboles de clasificación, para la variable respuesta *tipo de accidente* en función de factores concurrentes con el accidente, como son: tipo de vía, lugar de ocurrencia del accidente: recta, curva suave, etc., anchura de la calzada, anchura del carril, características del arcén y factores atmosféricos, permiten identificar algunos patrones de comportamiento específicos según los tipos más frecuentes de accidentes en los tramos de la muestra y los distintos factores de influencia sobre ellos. Las características de la vía (curvatura, ancho de calzada, ancho de arcén, ancho de carril) y meteorológicas condicionan el tipo de accidente.

1. INTRODUCCIÓN

1.1 Modelos de minería de datos aplicados al análisis de accidentes.

Los modelos no paramétricos de árboles de regresión y clasificación del campo de la minería de datos pueden ser muy útiles para analizar la relación entre la frecuencia o tasa o la severidad del accidente y un gran número de variables relacionadas con el tráfico, la infraestructura y factores concurrentes del accidente. En el campo de la accidentología se pueden citar como antecedentes a Lau y May, (1988, 1989), Karlaftis y Golías, (2002), Park y Saccomanno (2005), Forkenbrock y Hanley (2003), Chang y Chen (2005), Chang y Wang (2006), Sohn y Shin (2001), Sohn y Lee (2003), Abdel-Aty y otros (2005) y Kuhnert y otros (2000).

2.2 Datos para el desarrollo de modelos de árboles.

La base de datos de tramos del año 2007, contiene información proveniente de la base General de Accidentes (BGA) de la Dirección General de Tráfico y la base de datos del mapa de tráfico del Ministerio de Fomento. La misma contiene información seleccionada de 3.328 tramos de la red estatal de carreteras española (RCE y por tanto, excluye los tramos bajo otra jurisdicción: gobierno autonómico o de diputaciones) de cada una de las siguientes variables: número total de accidentes con víctimas ($A_T = 14.490$), la intensidad media diaria anual total de vehículos IMDT y los valores por tipo de vehículo: ligeros, motos, pesados y autocares ($IMDT_l$, $IMDT_m$, $IMDT_p$, $IMDT_b$ respectivamente), así como la velocidad media de vehículos ligeros (v_{ml}) obtenida en las estaciones permanentes de aforo de la red estatal. En la Tabla 1 se muestra un resumen de los valores estadísticos de algunas variables disponibles de los tramos, como la longitud, la unidad de exposición ($vk = 365.\text{long.}\frac{IMDT}{10^6}$) y variables del tráfico, de las que se pueden derivar otras de interés como es el porcentaje de vehículos pesados y de otros vehículos ($\%pes, \%liger$) como medida de la heterogeneidad y composición del tráfico en los tramos seleccionados.

		Media	95% Intervalo de confianza para la media		Mediana	DS	Mínimo	Máximo	Rango
			LI	LS					
$long$	AP	8,28	7,39	9,17	7,14	6,17	0,28	29,32	29,04
	AV	6,58	6,22	6,88	5,06	5,24	0,01	32,99	32,99
	DC	4,48	3,59	5,37	2,50	4,82	0,05	21,15	21,10
	C	6,93	6,66	7,19	5,00	6,17	0,10	47,22	47,12
	GLOBAL	6,81	6,61	7,01	5,01	5,89	01	47,22	47,22
v_{ml}	AP	121,91	119,71	124,10	125,00	15,22	92,00	174,00	82,00
	AV	115,53	114,48	116,58	118,00	16,58	40,00	175,00	135,00
	DC	79,16	75,81	82,51	79,00	18,13	39,00	110,00	71,00
	C	85,66	84,98	86,34	86,00	15,76	28,00	143,00	115,00
	GLOBAL	96,07	95,32	96,82	93,00	21,94	28,00	175,00	147,00

		Media	95% Intervalo de confianza para la media		Mediana	DS	Mínimo	Máximo	Rango
			LI	LS					
IMDT	AP	28409,45	24142,82	32676,08	19554,00	29574,851	360	149332	148972
	AV	23833,41	22198,39	25468,42	16970,00	25787,404	86	194482	194396
	DC	19290,19	16177,12	22403,26	15432,00	16852,120	426	111019	110593
	C	13375,10	12426,54	14323,65	6278,50	21995,61	22	198499	198477
	GLOBAL	17434,81	16616,70	18252,91	9555,50	24071,03	22	198499	198477
IMDPES	AP	4253,83	3401,12	5106,54	1996,00	5910,69	47	30474	30427
	AV	3714,94	3476,67	3953,21	2554,00	3757,99	14	29036	29022
	DC	2598,50	1960,53	3236,48	1789,00	3453,56	11	24634	24623
	C	1722,86	1612,63	1833,09	783,00	2556,10	4	26170	26166
	GLOBAL	2468,78	2353,74	2583,81	1338,00	3384,68	4	30474	30470
%pes	AP	13,79	12,63	14,95	12,21	8,03	2,17	50,05	47,89
	AV	17,66	17,05	18,28	15,39	9,69	1,73	95,61	93,88
	DC	14,41	12,39	16,43	11,65	10,93	1,42	49,12	47,7
	C	15,04	14,62	15,46	12,54	9,76	1,96	76,59	74,62
	GLOBAL	15,7	15,37	16,04	13,18	9,77	1,42	95,61	94,19
IMDLIG	AP	23876,94	20350,17	27403,7	16330	24446,37	311	127172	126861
	AV	19774,27	18332,5	21216,04	13684	22739,52	9	183867	183858
	DC	16334,54	13535,24	19133,83	12196	15153,57	386	106993	106607
	C	11469,97	10620,04	12319,89	5148	19708,43	18	174022	174004
	GLOBAL	14725,69	14004,67	15446,71	7666,5	21214,63	9	183867	183858
%liger	AP	84,95	83,82	86,09	85,97	7,86	49,6	96,9	47,3
	AV	81,08	80,47	81,69	83,41	9,65	2,2	97,32	95,12
	DC	84,17	82,24	86,11	87,32	10,47	50,33	97,55	47,22
	C	83,73	83,31	84,15	86,43	9,75	23,02	97,78	74,76
	GLOBAL	83,05	82,72	83,39	85,79	9,73	2,2	97,78	95,59
vk	AP	72,95	61,56	84,35	1,9	78,98	0,87	423,54	422,66
	AV	50,61	46,63	54,6	28,61	62,86	0	491,67	491,67
	DC	26,65	18,32	34,99	14,57	45,12	0,16	326,07	325,9
	C	24,66	22,93	26,39	11,35	40,12	0,01	369,99	369,98
	GLOBAL	34,91	33,12	36,7	16,29	52,68	0	491,67	491,67

Tabla 1 – Estadísticos descriptivos de las variables de tramo. Año 2007.

En cada uno de los tramos seleccionados, se han volcado los accidentes ocurridos y la información de algunos de los factores concurrentes de tipo categórico que se resumen en la Tabla 2.

Localización	Describe la geometría de la vía en el lugar donde tiene lugar el accidente	1. Recta 2. Curva suave 3. Curva fuerte sin señalizar 4. Curva fuerte con señal y sin velocidad señalizada 5. Curva fuerte con señal y velocidad señalizada
Tipo de vía	Especifica el tipo de vía o segmento	1. Autopista 2. Autovía 3. Vía de doble calzada 4. Vía convencional

Luminosidad	Describe las condiciones de luminosidad en el momento del accidente	1. Pleno día 2. Crepúsculo 3. Iluminación suficiente 4. Iluminación insuficiente 5. Sin iluminación
Superficie	Describe el estado de la calzada en el momento del accidente	1. Seca y limpia 2. Umbría 3. Mojada 4. Helada 5. Nevada 6. Barrillo 7. Gravilla suelta 8. Aceite 9. Otro tipo
Anchura de la calzada	Describe ancho de la calzada en 3 categorías	1. Menos de 5,99 m. 2. Entre 6 y 6,99 m. 3. De 7 m. o más
Anchura de carril	Describe el ancho del carril en el que se produjo el accidente	1. De más de 3,75 m. 2. De 3,25m. a 3,75 m. 3. Menos de 3,25 m.
Arcén	Describe las características (ancho) del arcén.	1. Inexistente o impracticable. 2. Menor de 1,50 m. 3. De 1,50 m. a 2,49 m. 4. De 2,50m. en adelante
Factores atmosféricos	Describe las condiciones atmosféricas del momento del accidente	1. Buen tiempo. 2. Niebla intensa. 3. Niebla ligera. 4. Lloviznando 5. Lluvia fuerte. 6. Granizando. 7. Nevando. 8. Viento fuerte. 9. Otro.

Tabla 2 –Variables concurrentes del accidente. Año 2007.

Un resumen del número de accidentes y reparto por tipos en cada tipo de vía se muestran en la Tabla 3.

	COLISIONES	SALIDAS	VUELCOS	ATROPELLOS	RESTO
AP	586	405	32	22	55
	8%	8%	7%	5%	10%
AV	2348	2086	153	111	164
	30%	39%	33%	26%	31%
DC	281	153	21	18	17
	4%	3%	5%	4%	3%
C	4509	2695	256	277	301
	58%	50%	55%	65%	56%
TOTAL TIPO DE ACCIDENTE	7724	5339	462	428	537
	53%	37%	3%	3%	4%
				TOTAL ACCIDENTES () 2007	14.490

Tabla 3 –Datos de los accidentes de la base de datos. Año 2007.

Las colisiones y salidas de la vía representan el 90% del total de accidentes (14.490) de la muestra seleccionada para el año 2007. Por tipo de vía, las colisiones se producen en un 58% en las vías convencionales, mientras que las salidas se reparten por igual entre vías de alta capacidad y convencionales. Los vuelcos en vía convencional representan el 55% del

total. Las características de la vía o del tramo influyen en el tipo de accidente y determinan patrones de comportamiento que no emergen del análisis estadístico pero que se pueden determinar a través de modelos de árboles.

2. ÁRBOLES DE CLASIFICACIÓN. PATRONES DE COMPORTAMIENTO DE ACCIDENTES Y FACTORES CONCURRENTES.

En la Figura 1 se muestra el árbol CART (Breiman et. al. 1984) para la variable dependiente tipo de accidente (2 tipos: colisiones y salidas) en función de las variables predictoras: tipo de vía, ancho de arcén, ancho de la calzada, ancho de carril y factores atmosféricos (recodificada en dos niveles: buen tiempo=1 y resto=2).

El criterio de división utilizado para el crecimiento del árbol es el de clasificación errónea o "towing index", como medida del grado de impureza, determinado mediante $I_M(\mathrm{m}) = \frac{\sum_{l=1}^{n_m} 1(y_{lm}, y_k)}{n_m} = 1 - \pi_k$, donde y_k es la categoría o nivel modal del nodo con probabilidad ajustada π_k, la función $1()$ es una función indicador, que toma el valor 1 si $y_{lm} = y_k$ o 0 en caso contrario. Una medida de bondad de ajuste del modelo se puede obtener a través de la tasa de clasificación errónea / correcta o medida de impureza, obtenida como el porcentaje de observaciones clasificadas en un nivel diferente de los valores observados. Giudici, P. (2003).

Las variables seleccionadas en el modelo por orden de importancia normalizada son: el lugar de ocurrencia del accidente (100%), seguida del ancho de arcén (11,1%), tipo de vía (10,9%), el ancho de la calzada (6,5%), los factores atmosféricos concurrentes (4,8%) y el ancho de carril en último lugar (4,5%). (Figura 2).

En la Tabla 4, se muestra la tabla de clasificación (confusion matrix) del modelo, de la que se obtiene el porcentaje de fallos en la clasificación o porcentajes de clasificaciones erróneas de observaciones en un nivel distinto del valor observado (índice de clasificación errónea), para una medida global de la precisión del modelo que alcanza el 67,2% (8.774 clasificaciones correctas de 13.063), siendo ésta más alta para colisiones que para salidas.

El coste de la clasificación incorrecta (o tipo de error en la literatura) de los dos tipos de accidentes [error tipo I o clasificación errónea de colisiones= 2.975/ 9.385 y error tipo II o clasificación errónea de salidas= 1.314/3.678] es un criterio a ser tenido en cuenta para la elección de un modelo entre varias opciones, considerando paralelamente el coste del aumento de complejidad (o alejamiento del principio de parsimonia) del mismo. La elección del mejor modelo de entre los que compiten, depende de la información o el conocimiento del coste de los errores, siendo aconsejable en cualquier caso, el modelo CART cuando esta información se desconoce o el coste de ambos tipos de error es equivalente. En nuestro caso, un criterio de interés podría ser minimizar el tipo de error I frente al error tipo II, por la severidad que tienen las colisiones frente a las salidas de la vía.

Porcentaje de Clasificación

Observado	Pronosticado		
	1	4	Porcentaje correcto
1	6410	1314	83,0%
4	2975	2364	44,3%
Porcentaje global	71,8%	28,2%	67,2%

Variable dependiente: 1=COLISIONES, 2=ATROPELLOS, 3=VUELCOS, 4=SALIDAS DE VIA, 5=RESTO ACC

Tabla 4 –Matriz de clasificación del modelo CART para colisiones y salidas. Año 2007.

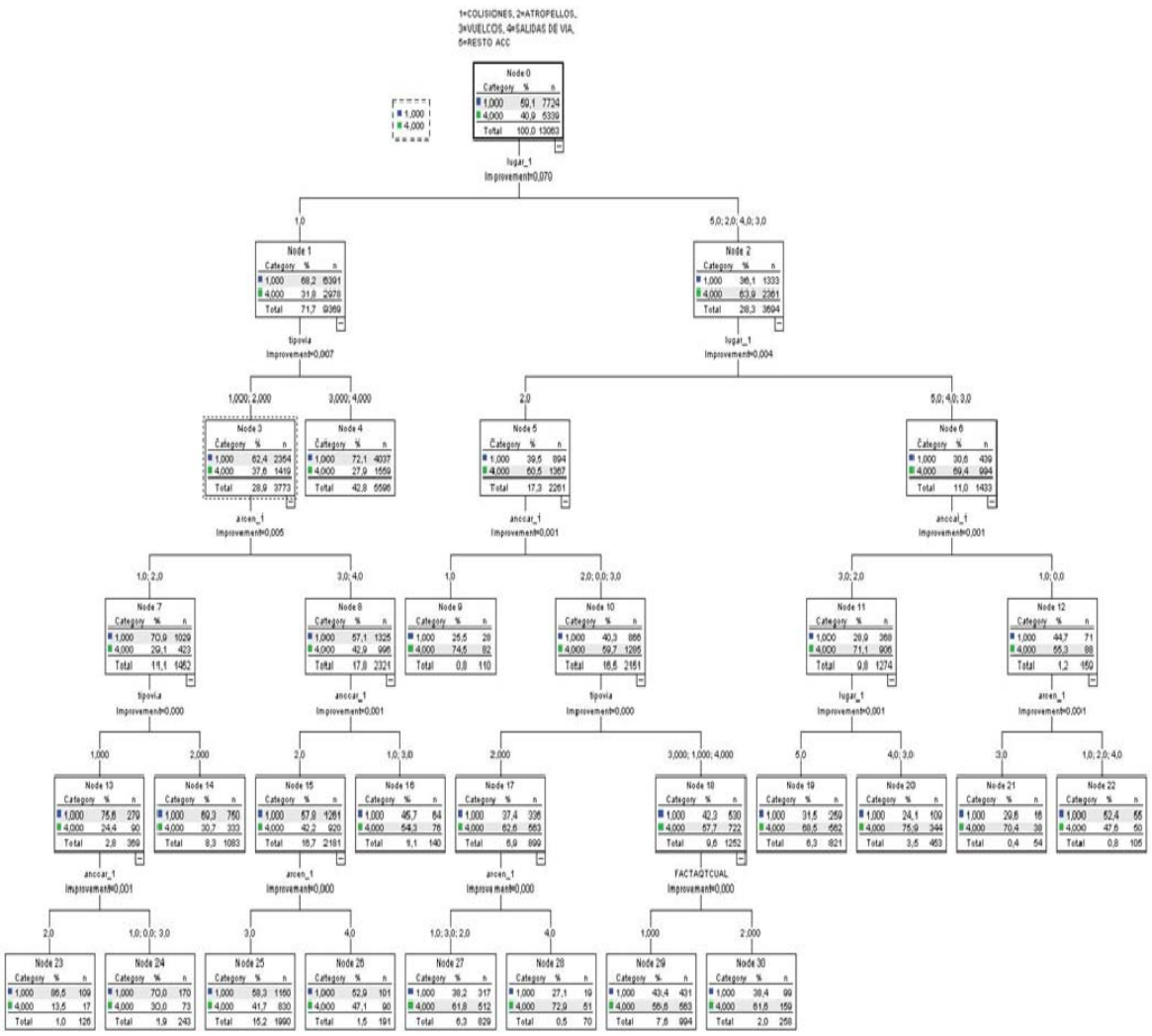

Figura 1 –Modelo de árbol de clasificación para el tipo de accidente. Año 2007.

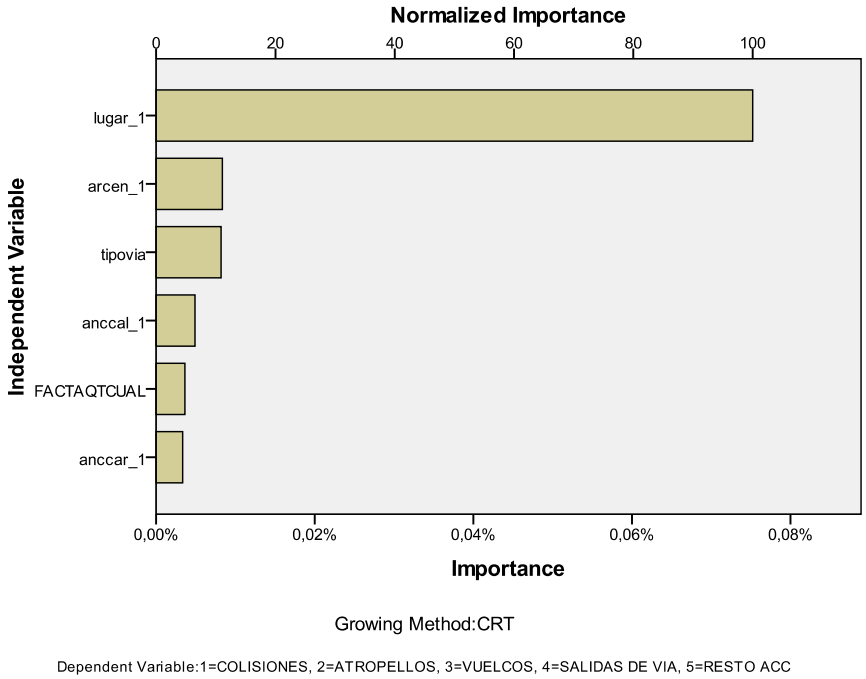

Figura 2 –Importancia de las variables del modelo de clasificación para colisiones y salidas de la calzada. Año 2007.

Para la interpretación de los patrones de comportamiento del árbol de clasificación (Figura 1), se ha utilizado la regla discriminante más común: la regla de la mayoría (*majority rule*) por la cual, todas las observaciones clasificadas en el nodo corresponden al nivel o modo más frecuente. Con esta regla los patrones que se destacan son:

- El lugar de ocurrencia del accidente discrimina en primer lugar la ocurrencia de un tipo de accidente u otro: en recta se producen principalmente colisiones, mientras que las salidas de la calzada se producen en curva (suave y fuerte con o sin señalización).
- En los tramos rectos en los que se producen fundamentalmente colisiones, adquiere importancia el tipo de vía y se distinguen claramente las colisiones en vías de alta capacidad de las que se producen en vías convencionales. (Junto a estas resultan seleccionadas las vías de doble calzada, de los que hay 115 tramos frente a 2068 de convencionales).
- Las colisiones ocurridas en secciones recta y fundamentalmente en vías convencionales, determinan escenarios concretos en los que las características de la sección de carretera cobran importancia: ancho de la calzada, ancho de carril e inexistencia de arcén (o insuficiente) para la realización de maniobras de evitación de la colisión.
- Las salidas en curva, evidencian un patrón de comportamiento diferente según se produzcan en curvas suaves o fuertes. La ocurrencia de salidas se incrementa en secciones curvas fuertes con o sin señalización de velocidad.

- En los accidentes por salida de la calzada en secciones de curva suave, adquiere importancia el ancho de carril y tipo de vía de modo análogo a lo detectado en secciones rectas.
- Las salidas se incrementan bajo condiciones atmosféricas adversas (lluvia intensa, niebla, viento fuerte). Bajo estas condiciones adquieren importancia factores como la adherencia y la visibilidad, entre otros.

3. ÁRBOLES DE REGRESIÓN. PATRONES DE COMPORTAMIENTO DE ACCIDENTES Y VARIABLES DE TRÁFICO.

En la Figura 3 se muestra el modelo de árbol de regresión para el número total de accidentes en cada tramo (A_{Ti}) en función de los predictores: tráfico total, porcentaje de vehículos pesados, velocidad media de vehículos ligeros y tipo de vía. En la Figura 4, se muestran el orden de importancia normalizada de las variables seleccionadas en el modelo de regresión: la primera es la intensidad media diaria anual de tráfico cuya importancia normalizada se sitúa en el 100%, seguida del porcentaje de vehículos pesados (4 %), velocidad media de vehículos ligeros (2,7%) y tipo de vía (1,7%). La variable de mayor influencia para la explicación de la frecuencia de accidentes en las carreteras es el tráfico medio anual, medida indirecta de la exposición en un tramo concreto. A medida que aumenta el tráfico total, mayor es el número medio de accidentes. El segundo factor que explica el número de accidentes es el porcentaje de vehículos pesados y también aumenta con el mismo. Por tipo de vía, hay mayor número de accidentes en autopistas y vías convencionales. Adicionalmente se han ajustado modelos para los dos principales tipos de accidentes: colisiones (COL_i) y salidas (SAL_i) y se calculado el R^2 como medida de bondad de ajuste de los 3 modelos: $R^2[A_{Ti}] = 29\%;\ R^2[COL_i] = 33\%;\ R^2[SAL_i] = 16\%$, que pone de manifiesto la superioridad del modelo de colisiones. El orden de magnitud del R^2 obtenido es similar al que se ha obtenido con modelos lineales generalizados para la misma variable dependiente. (Arenas R. B. et. al. 2009). La selección de variables por orden de importancia de los modelos de regresión para colisiones y salidas se muestra en la Tabla 5.

COLISIONES: orden de importancia de las variables independientes

Variable Independiente	Importancia	Importancia Normalizada
IMDTRA	6,446	100,0%
VMEDIA	0,533	8,3%
TIPOVIA	0,202	3,1%
$\%pes$	0,108	1,7%

SALIDAS: orden de importancia de las variables independientes

Variable Independiente	Importancia	Importancia Normalizada
IMDTRA	0,769	100,0%
$\%pes$	0,282	36,6%
VMEDIA	0,208	27,1%
TIPOVIA	0,090	11,7%

Tabla 5 –Importancia de las variables del modelo de regresión para el número de colisiones y salidas. Año 2007.

En las colisiones la variable más relevante es la intensidad media diaria anual, mientras

que en las salidas adquieren importancia otras variables (el orden de magnitud de $\%pes$ y velocidad media son comparable entre sí y se acercan a la de la intensidad de tráfico), para la explicación de la frecuencia de estos accidentes. Las salidas en secciones curvas, por pérdida de control del vehículo pueden estar relacionadas con la velocidad y depende del tipo de vía, del ancho de la calzada, etc. Este patrón de comportamiento se puso de manifiesto a partir los resultados de los árboles de clasificación.

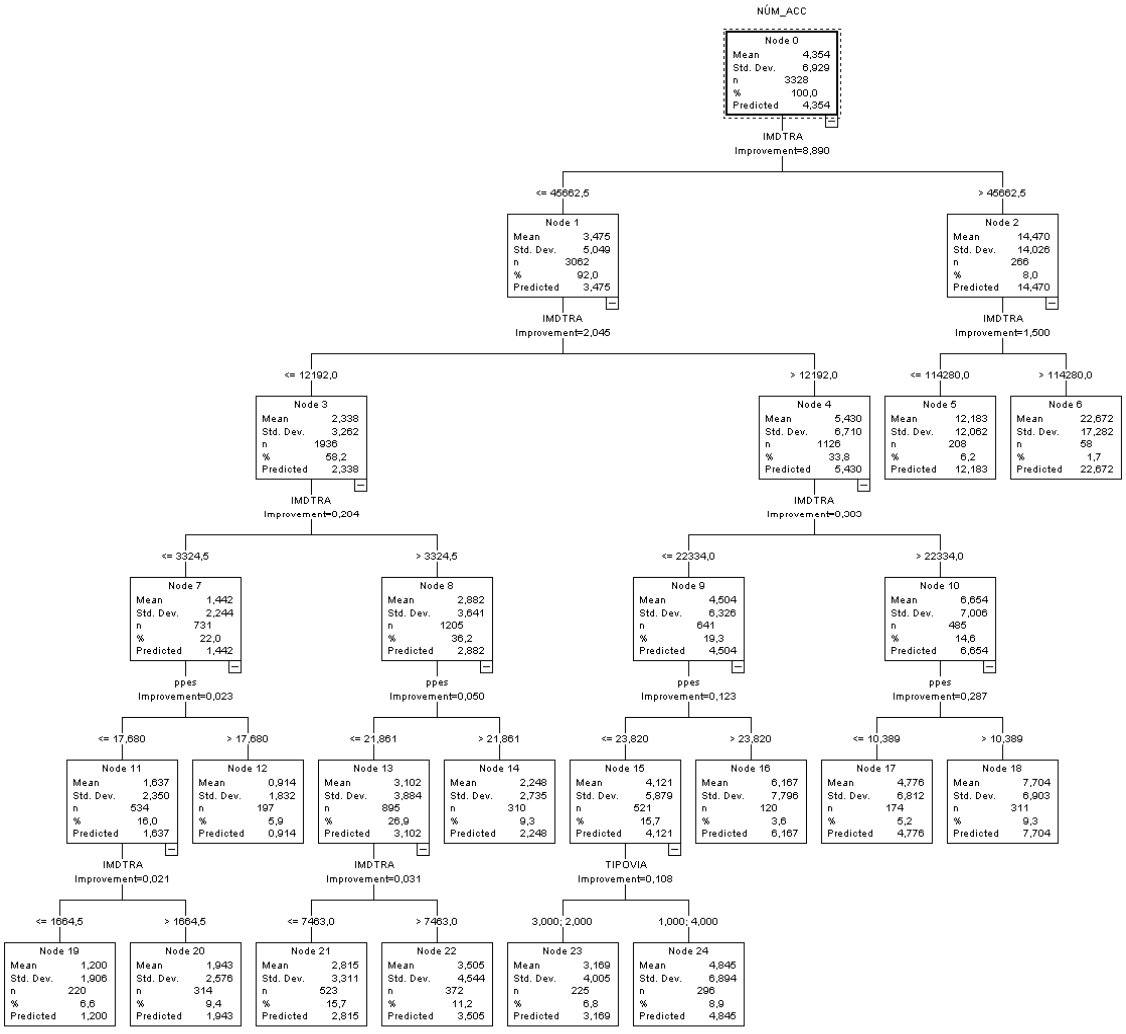

Figura 3 –Modelo de regresión para el número total de accidentes. Año 2007.

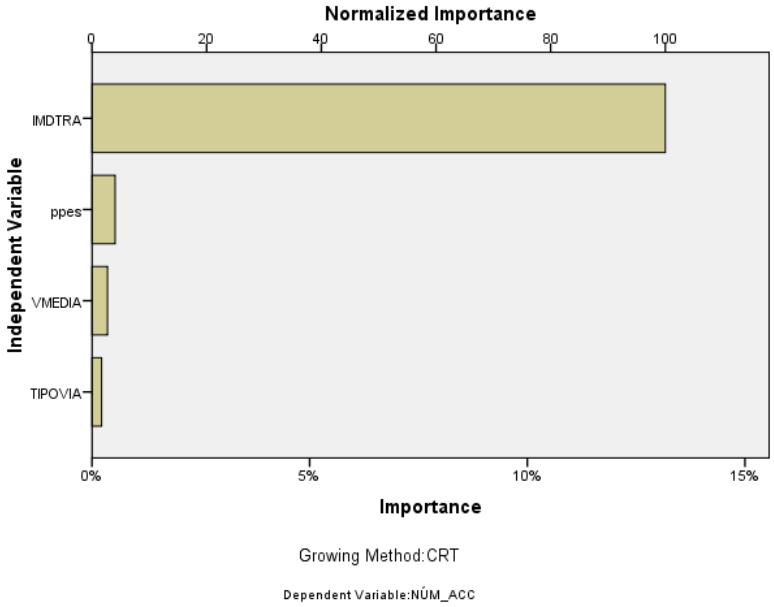

Figura 4 – Importancia de las variables del modelo de regresión para el número total de accidentes. Año 2007.

4. CONCLUSIONES

Los modelos desarrollados constituyen una alternativa a los modelos paramétricos por su capacidad de identificar patrones a partir de datos, sin la necesidad de establecer una relación funcional entre la respuesta y los factores de influencia. A través de los modelos de clasificación se pueden determinar patrones o interacciones entre variables que no son posibles de establecer de manera directa, mediante las técnicas comunes de exploración de datos y de modelado estadístico. Los modelos de árboles de regresión son herramientas útiles y poderosas para evaluar la preminencia de predictores y las interacciones entre las variables disponibles.

5. REFERENCIAS

- Abdel-Aty M., Keller, J., Brady, P.A. (2005). Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transportation Research Record 1908. 37-45.
- Arenas Ramírez, B., Aparicio Izquierdo F., González Fernández C., Gómez Méndez, A. (2009). The influence of heavy goods vehicle traffic on accidents on different types of Spanish interurban roads. Accident Analysis and Prevention 41: 15-24.
- Base de datos de accidentes, 2007. Dirección General de Tráfico.
- Breiman L.; Friedman J.H.; Olshen, R.A.; Stone, C.J. (1994). Classification and Regression Trees. Wadsworth and Brooks / Cole, Monterrey.
- Chang L-Y., Chen W-C. (2005). Data mining of tree-based models to analize freeway accident frequency. Journal of Safety Research 36. 365-375.
- Chang L-Y., Wang H-W. (2006). Analysis of traffic injury severity: An application

of non-parametric classification tree techniques. Accident Analysis and Prevention 38. 1019-1027.

- Forkenbrock D.J., Hanley P.F. (2003). Fatal crash involvement by multiple-trailer trucks. Transportation Research Part A 37. 419-433.
- Guidici P. (2003). Applied Data Mining. Statistical Methods for Business and Industry. John Willey& Sons Ltd, The Atrium, Southern Gate. Chichester. West Sussex PO19 8SQ. England.
- Karlaftis M.G., Golias I. (2002). Effect of road geometry and traffic volumes on rural roadway accidents. Accident Analysis and Prevention 34(3). 357-365.
- Kuhnert P.M., Do, K.A., McCluer, R. (2000). Combining non-parametric models with logistic regression. An application to motor vehicle injury data. Computational Statistical Data Analysis 34(3). 371-386.
- Lau M.Y.K, May A.D. (1988). Accident Prediction Model Development: Signalized Intersections. Research Report UCB-ITS-RR-88-7. Institute of Transportation Studies. University of California. Berkeley. Ca.
- Lau M.Y.K, May A.D. (1989). Accident Prediction Model Development for Unsignalized Intersections. Research Report UCB-ITS-RR-89-12. Institute of Transportation Studies. University of California. Berkeley. Ca.
- Mapa de tráfico, 2007. Ministerio de Fomento. Dirección General de Carreteras. Madrid.
- Park Y.J., Saccomanno, F. (2005). Collision frequency analysis using tree-based stratification. Transportation Research Record 1908. 121-129.
- Sohn S.Y., Lee, S.H. (2003). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Safety Science. 41(1). 1-14.
- Sohn S.Y., Shin, H. (2001). Pattern recognition for road traffic severity in Korea. Ergonomics 44(1). 107-117.